图书情报工作
第62卷 第3期 2018年2月

面向科技文献知识表示的知识元本体模型*

秦春秀 杨智娟 赵捧未 刘杰
西安电子科技大学经济与管理学院 西安 710071

摘要:[**目的/意义**]随着科技文献资源的急剧增长,用户淹没在科技文献的海洋中,为用户提供快速、精准的细粒度知识元服务将成为未来文献知识检索的发展趋势。[**方法/过程**]在分析科技文献文本结构的基础上,逐步深入到科技文献的内容中,以期通过构建一种面向科技文献知识表示的知识元本体模型,将科技文献内容中句义完整的细粒度知识点表示成具有统一结构的知识元。[**结果/结论**]以一篇科技文献为实例,展示笔者提出的基于知识元本体模型的科技文献知识表示方法,但该示例仅呈现了科技文献中引言部分的相关知识点,需进一步验证该知识元本体模型的有效性。

关键词:科技文献 知识表示 知识元 本体模型
分类号:G302
DOI:10.13266/j.issn.0252-3116.2018.03.012

科技文献是科技工作者及科研人员进行科学探索的结晶,蕴含着丰富的专业化科技成果和重要的科研知识发现,极具参考价值和应用价值。科技文献知识资源的共享与交流不仅有利于挖掘科技文献的潜在价值,更能促进知识的学习、创新与发现,推动科学的发展与进步。在大数据时代下,科技文献不仅总量巨大,而且各领域科技文献的学术期刊以平均每年4.7%的频率呈爆炸式增长,科技工作者淹没在海量科技文献的海洋中[1]。

现今人们汲取所需的科技文献知识资源时,通常借助传统的信息检索技术,获取以篇章为单位的科技文献,逐篇浏览文献中蕴含的知识点,人工筛选出所需的相关知识点,这极大地耗费了科研人员的时间及精力,降低了科技文献知识资源获取的效率。产生这样低效的文献知识获取方式的根源在于,当前的检索工具通常以粗粒度的文献为基本单元进行文献的描述、表示及组织,而未将文献的知识控制单元细化到知识点,导致信息检索系统难以快速、精准地匹配到用户真正所需的知识点,造成了资源海量、知识饥渴的窘状。

为了解决上述问题,目前,学者们主要从知识点的特征和科技文献的知识结构两个角度进行深入研究。①从某类知识点的特征着手,构建该类知识点的抽取规则,并依照规则匹配算法,从科技文献文本内容中抽取出定义[2]、方法[3]或创新点[4]等单一类型的知识点,而且这些知识点一般分散在文献的各章节区域中,较难全面地定位到所有相关的知识点,使得知识点抽取的效率及精准度降低;②为了更充分地挖掘科技文献中的知识资源,尝试通过对科技文献中某一区域或模块的知识结构进行分析,给出一个资源描述框架[5],用于挖掘文献正文区域中所包含的承载着完整科学研究思路的知识点集合,如研究的领域、背景、问题、理论、方法、评价等知识点,但该方法缺乏对科技文献外部特征的描述,无法详尽地描述出科技文献内容中所包含的各类知识点和知识结构,未能统一表示出这些知识点内部的属性结构,不利于知识的推理与发现。

因此,笔者尝试从文献知识组织的基础环节——知识描述与表示环节,通过系统分析科技文献文本结构,以知识元本体理论为基础,研究科技文献的细粒度知识表示方法,以期从细粒度视角给出一种全面、规范的科技文献知识表示法,旨在为统一描述科技文献知识结构及其内所含的学术知识点,实现细粒度的精准知识点服务提供理论支持。

* 本文系国家自然科学基金项目“知识社区中的资源语义空间及其检索研究”(项目编号:71573199)研究成果之一。
作者简介:秦春秀(ORCID:0000-0001-9524-9359),副教授,硕士生导师,E-mail:qinchx@126.com;杨智娟(ORCID:0000-0001-6678-7983),硕士研究生;赵捧未,教授,博士,博士生导师;刘杰(ORCID:0000-0002-6730-8056),硕士研究生。
收稿日期:2017-08-23 修回日期:2017-11-15 本文起止页码:94-103 本文责任编辑:王善军

1 文献知识表示研究现状

文献知识表示是文献知识组织、检索和应用的基础,其完备性及适用性影响着文献知识服务的水平。目前,根据研究者们是否对知识对象的内部结构关联进行描述,将文献知识表示方法分为两类,即基于特征的知识表示方法[6]和结构化的知识表示方法[7]。

基于特征的知识表示方法将知识对象视为相互独立的个体,通常以“属性－值”对、n元组、特征向量等形式,清晰地描述出知识对象的属性特征,便于对知识对象进行识别及标引。该类方法中具有代表性的方法有:①纳米出版物模式,以简易的RDF三元组生成命名图的形式,描述科学文献的结论出处、结论的背景及语境信息等[8],将文献中的科学结论表示成具有统一结构的最小出版信息单元,却忽略了这些信息单元之间存在的潜在科学规律;②知识元表示方法,以n元组的形式,描述期刊文献知识点的编号、来源、分类、内容等特征[9],将文献中细粒度的知识点表示成内容相对完整且结构一致的知识元,这些知识元则是在一定的语境下语义相对完整的知识单元,其形式上具有多样性,内容上具有相对独立性和完整性[10]。当知识元的内容增加到一定程度时,知识元可分解成更小粒度的知识元,故“知识元应具备内容小粒度、可链接成知识网络、可进化发展、可重构、支持语义检索等功能”[11]。然而,当描述科技文献中一个具体的论证性过程时,该知识元表示法仅能细粒度化地描述出一个个论点、论据、论证方法及结论等相对独立的知识元,却未能清晰地表示出该文献依托这些知识元集合所构筑的一个语义完整的论证关系链。

为了详细描述文献内知识对象之间复杂的语义关系,学者们给出了结构化的知识表示方法,如基于图结构的表示法[6]、面向对象表示法[12]、产生式框架表示法[13]、本体表示法[14]等。①基于图结构的表示法,将文献中的知识点表示成一个不可再分割的知识或概念节点,由节点和节点之间的语义边共同搭建成一个图式结构,展现文献内知识点之间的语义关联,代表性的方法有知识网络[7]、语义网络[15]、概念图[16]等知识表示法。该类方法在对文献知识点之间语义关系的表示方面具有较强的知识表现力,但无法全面描述出文献知识或概念节点的内部特征,且随着知识或概念节点数呈指数倍增长,文献知识点的检索难度及成本也随之大大增加。②面向对象表示法,将知识对象的属性、方法等封装到结构化的模块中,实现了对象间的继承与演化,用于描述文献内容中基本概念等,但缺少对知识对象间继承之外的其他语义关系进行动态表示及推理。③产生式框架表示法,以填充槽的框架形式,描述文献内知识对象的静态知识结构,以产生式规则表示知识对象之间的推理规则,但不适合复杂的推理关系。例如,采用产生式框架表示法来描述《说文解字》中静态知识结构及其间的音义关系,当知识对象之间的推理关系复杂时,会导致其推导效率降低。④本体表示法,拥有强大的逻辑推理能力,以概念、属性、关系、实例等形式,描述各主题领域内文献资源中知识对象之间的语义逻辑关系,完成知识对象之间的动态推理,如“江海文化”知识本体[14]、专利本体[17]等。然而文献的主题分类较多,表示不同主题的本体缺乏统一的结构,不利于知识的表示、共享及交换。

上述方法中,基于特征的知识表示法中知识元表示方法能详细描述科技文献内知识点的属性特征,却未能全面表示出知识点之间系统性的语义关系;而用于梳理知识点间复杂语义关系的结构化知识表示方法中,本体表示法能标准化地描述出不同主题领域内科技文献知识点之间的系统性语义逻辑关系,却难以突破领域性限制,统一表示科技文献内知识点的结构。

知识元本体构建定义了知识元的组织骨架模型,建立了知识元与本体元素之间的联系,以概念、属性、方法、关系四元组形式,全面地描述了知识对象中语义内容的内在结构和语义关联,揭示了知识元的属性特征和知识元之间的语义逻辑关系,是对领域知识的规范化抽象及描述,便于对知识进行语义推理,实现知识的组织与检索[18-19]。知识元本体已广泛应用于wiki知识元标注[18]、学科知识的语义标引[11]、“粤海关”文献知识元的归类及其间的组合链接[20]、饮食与疾病领域知识元之间的逻辑推理[21]、基于文献主题成因的知识发现[22]等场景中。因此,笔者以知识元本体理论为基础,吸纳知识元及本体表示法的思想,给出一种面向科技文献知识表示的知识元本体模型,以期解决由科技文献粗粒度的知识表示方法导致的知识获取低效性问题。

2 科技文献文本结构分析

一篇科技文献通常具有物理结构和逻辑结构,物理结构呈现出文献的主观认知结构,即标题、作者、机构、章、节、段、句、词、引文;在此基础上,对文本进行层次划分,挖掘文本中不同层次下内容的主题,得到文本的逻辑结构,即篇章主题、层次主题、段落主题、句子主

题、主题词、标识词及分类号，以期表示出文本内容中的知识结构[23]。仅以单一的主题分类法描述文献内容的知识结构，忽略了文献中简单、直观、少有歧义的非主题文献特征，导致无法在文献检索过程中准确排除大量不相关的文献[24]。故整合上述文献特征，将科技文献的特征统一划分为外部特征和内容特征，外部特征一般包括符号标识、书名、作者、机构名、文献分类等，内容特征包括主题词、分类号等[25]，便于清晰、有效、全面地描述出科技文献中细粒度知识点的特征。

在实际应用中，科研文献本身具有由标题、作者、摘要、关键词、正文构成的特定文本结构[26]；逐层深入到正文内容中，呈现出一种通用结构，即引言、方法、结果、讨论[27]。这种通用结构在科技文献内容的表达上具有一定的章节性语义划分功能。Y. Ding 等采用文献分析法，对该通用结构进行细化，形成由摘要、引言、相关研究、方法、实验/结果、结论构成粗粒度的功能结构[28]。王鹏等依照分层分割的文本处理方法[29]，逐层细分上述功能结构中每部分的知识块，得到由科技文献内细粒度知识点构成的层次性功能结构。在科技文献的摘要部分，通常概述了科学研究的背景、结论、方法、对象及结果，尤其是对研究结论、方法和结果的概述最为频繁[30]。在科技文献正文中，引言包含背景知识、问题分析、主要工作三类信息[31]，而在问题分析之前需要引入研究问题，描述研究的由来、研究动机、研究目的等内容[32]；方法分为科学研究方法和问题解决方法，其中科学研究方法包括问卷调查、专家访谈、案例分析法等[3]，问题解决方法按表现形式分为模型、算法和指标，其中模型包括框架类模型和数学模型[33]；实验主要论述实验数据、实验过程、实验结果、评测、实验发现及讨论，其中，实验过程包括系统的设计、实现等[32]，另外，经文献调研表明，在评价的实践中主要对评价指标、评价方法、评价程序等问题进行研究[34]；结论主要包括突出贡献的阐述、（非）预期结果的说明、结果的推广、未来的研究方向等内容[35]。

上述科技文献中的层次性功能结构，将科技文献中的知识资源分割为细粒度的功能性知识点，这种层次性功能结构满足人们对科技文献知识点的使用需求，便于快速了解研究背景、紧跟学术前沿等。由科技文献的文本结构，构建出科技文献的分层信息模型，能够对科研热点的发现、科研内容的相似性和分类比较提供更加准确的基础信息。因此，笔者将深入到科技文献的文本内容中，挖掘科技文献内容中的层次性功能结构，旨在构建一个用于表示科技文献知识的知识元本体模型，统一、全面地描述科技文献中细粒度的功能性知识点特征及其文本结构中知识点之间的逻辑关系，为实现科技文献知识点的精准性检索服务提供理论基础。

3 科技文献的知识元本体模型

知识元本体结构通常被表示成四元组形式[18]，即 K = (C, P, M, R)，通过概念的属性集 P 及方法集 M 描述文献知识元的概念 C 特性，通过概念之间的语义关系集 R 描述知识元的内部体系结构及知识元之间的网络化结构，为知识元语义链接的构建提供保障。已有的知识元本体结构中的方法集与关系集仅分别表示出概念之间的函数关系及语义关系，未能充分描述概念的属性之间的函数关系以及概念、属性、方法三者之间的语义关系。

在对科技文献文本结构的系统性分析的基础上，笔者对上述知识元本体结构进行延伸及拓展，实现科技文献知识元本体模型的设计。借助本体中概念集的层次化结构，清晰地描述出科技文献中多层次的文本结构；而每一概念能抽象地描述出科技文献中每一类知识元，通过对概念的属性特征进行细致地刻画，形式化地表示科技文献内各类知识元的内部结构；定义本体模型的方法集，描述概念之间、概念与属性之间、属性与属性集之间所存在的函数关系，明确科技文献中知识元所属的概念、属性之间的语义规则，规范化表示出科技文献中知识元之间内在的语义关联；定义概念、属性、方法之间的关系集，系统描述科技文献中知识元之间的逻辑关系，精准呈现科技文献中隐含的科学研究思路。

因此，笔者将在已有的知识元本体理论的基础上，依照上述科技文献知识元本体的设计思路，通过分析科技文献的外部及内容特征，定义一个科技文献知识元本体模型及其内部的概念、属性、方法及关系四大构成要素，旨在为统一描述科技文献内部细粒度知识点特征及知识关联，提供一种面向科技文献的知识表示方法。

3.1 科技文献知识元本体模型的数学描述

一篇科技文献是由多个内容上相对独立且语义上相互关联的知识元构成的一个语义相对完整的小型知识库。为了便于用户精准、细粒度地检索到所需知识元，笔者从微观层面构建一种科技文献知识元本体模型，统一描述并表示科技文献知识元的组成及知识元之间的语义关系。该知识元本体模型可形式化表示如

下：

$$KEO_{SL} = \{C, P_C, M_{C,P}, R_{C,P,M}\}$$

其中，KEO_{SL}表示科技文献知识元本体，C 表示科技文献知识资源中概念集，P_C 表示概念 C 的属性集，$M_{C,P}$表示概念 C 及其属性 P 相关的方法集，$R_{C,P,M}$表示概念 C、属性 P、方法 M 之间的语义关系集。

对于任意的科技文献知识元 e，其本体结构中概念集 C_e、属性集 P_e、方法集 M_e、关系集 R_e 的结构及其之间的关系，用数学公式表示如下。

（1）概念 C_e 的属性集：

$$P_e = (PD_e, PO_e, D_P)$$

$$PD_e = (DC_e, V_e, D_{PD})$$

$$PO_e = (PF_e, T_e, D_{PO})$$

其中，D_P 表示概念的属性是否可描述或可测量，可分为 D_{PD}和 D_{PD}；概念的数据属性 PD_e 结构中 $DC_e = \varphi$ 且 DC_e 表示数据属性 PD_e 所属的定义域为概念集 C_e，V_e 表示数据属性 PD_e 的取值类型，D_{PD}表示数据属性 PD_e 是否可描述或可测量；概念的对象属性 PD_e 结构中 PF_e 表示该对象属性 PO$_e$ 所属的父属性，T_e 表示对象属性 PO_e 的特性，D_{PO}表示对象属性 PO_e 是否可描述或可测量。

（2）概念 C_e 及其属性 P_e 相关的方法集：

$$M_e = (MF_e, MOCG_e, D_M)$$

其中，MF_e 表示方法的公式描述；$MOCG_e$ 表示方法涉及的对象类型组合，包括概念及其属性组合、概念与概念的组合；D_M 表示方法是否可描述或可测量。

（3）概念 C_e、属性 P_e 及方法 M_e 之间的关系集：

$$R_e = (RFM_e, PO_e, RLM_e, ROCG_e, D_R)$$

其中，RFM_e 表示语义关联前者；PO_e 表示对象间的语义关系，即对象属性集；RLM_e 表示语义关联后者；$ROCG_e$ 表示语义关联对象类别组合，包括概念与概念组合、属性与属性组合、方法与方法组合、概念与属性组合、概念与方法组合、属性与方法组合；D_R 表示语义关系是否可描述或可测量。

D_X 表示数据属性 D_{PD}、对象属性 D_{PO}、方法 D_M 或关系 D_R 是否可描述或可测量，$D_X = 0$ 表示关系不可描述，$D_X = 1$ 表示关系可描述但不可测量，$D_X = 2$ 表示关系可描述且可测量。

由上述科技文献知识元本体结构 KEO_{SL}可知，科技文献知识元 e 的本体结构的组成元素可具体表示为概念集 $C_e = \{C_1, C_2, C_3, \cdots\}$、属性集 $P_e = \{P_1, P_2, P_3, \cdots\}$、方法集 $M_e = \{M_1, M_2, M_3, \cdots\}$ 和关系集 $R_e = \{R_1, R_2, R_3, \cdots\}$。该科技文献知识元本体结构，可清晰地描述出科技文献的知识元结构和知识元之间的语义关系。

3.2 科技文献知识元本体模型的要素定义

依据上述给出的科技文献知识元本体的形式化数学模型，定义出科技文献知识元的概念集、属性集、方法集及关系集的具体内容。同时，依据 T. R. Grubers 的本体设计原则[36]，对科技文献知识元本体模型的每个设计阶段进行规范，描述了科技文献中语义层面的知识，保证了本体模型具有明确性、客观性、一致性、可拓展性和最小本体承诺[37]。

3.2.1 科技文献知识元本体的概念集　在概念集的设计阶段，要保证本体模型具有明确性、客观性、一致性和可拓展性，明确各类概念术语的涵义，避免二义性，确保概念类具有语义一致性，且在添加新概念时无需修改已有内容，使得该知识元本体模型中的概念集能全面、抽象化地描述科技文献中不同层次下的各类知识元。

从科技文献特征角度，将一篇科技文献明确地划分为内容特征和外部特征两大概念类，对每一概念类进行扩展，获得该科技文献知识元本体的概念集。科技文献的外部特征包括符号标识、书名、正文语种、作者、出版时间、出版社、期刊、会议、文献类型九大类[25]，其中，科技文献的类型又包括科技图书、科技期刊、专利文献、会议文献、科技报告、政府出版物、学位论文、标准文献、产品资料和其他文献等[38]。科技文献的内部特征包括学科、主题、分类号、关键词、创新点以及摘要、引言、研究现状、核心研究内容、实验与评价、结论与展望等功能性特征。科技文献知识元本体的概念及概念间的层次结构见图 1。

3.2.2 科技文献知识元本体的属性集　在属性集的设计阶段，保证本体模型具有明确性、客观性、一致性和最小本体承诺，各类概念要具有相对统一、清晰、简洁的属性集，使得该知识元本体模型能统一描述出各类知识元的内部结构。故将科技文献知识元本体中概念类的属性分为两大类，即数据属性和对象属性，且同类属性之间存在一定的层次关系。数据属性描述概念类固有特性的数据特征，即科技文献知识元本体中概念类的属性集 P，概念子类继承其父类的数据属性；对象属性指概念类或属性之间的语义关系，具体描述科技文献知识元本体中的部分关系集 R。

在科技文献知识元本体中，数据属性包括内容特征类中概念的属性和外部特征类中概念类的属性。在内部特征类中，摘要、引言、研究现状、核心研究内容、

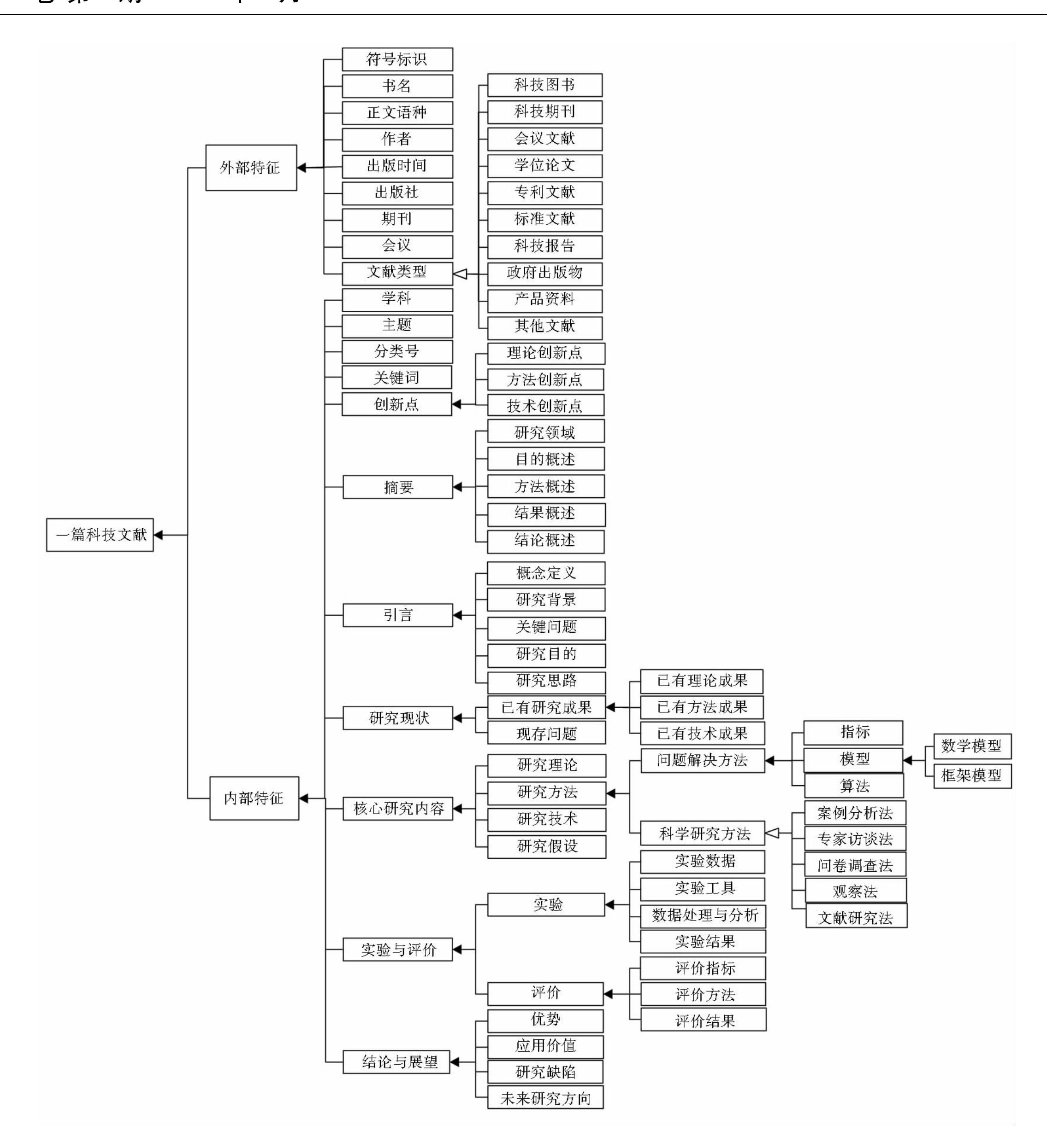


图 1　科技文献知识元本体的概念层次结构

注：矩形方框表示一个概念，实心箭头代表概念与概念之间的部分与整体关系，空心箭头代表概念与概念之间的继承关系

实验与评价、结论与展望类均具有知识标识号、知识元名称、知识导航、原文内容、文本长度、表示形式、特征词、引文编号等属性[9]；在外部特征类中，符号标识的属性有文献编号、专利号、标准号、报告号，书名的属性包括题名和篇名，出版社的属性有出版社名称、出版社地址，期刊的属性有期刊名称、卷/期/页、会议的属性有会议名称、会议地址，正文语种的属性有汉语、外语，作者类的属性包括编号、地址、单位、名字、性别、国籍、职称。

在科技文献知识元本体中，对象属性包括同位关系和等级关系两种[25]：同位关系可细分为创新关系、同义关系（如阐述关系）、定位关系、相关关系、顺序关系等，其中，相关关系包括交叉关系（如合作关系）、并列关系、互斥关系，顺序关系可细分为引出关系、次序关系、演化关系、蕴含关系、撰写关系等；等级关系可细分为部分与整体关系、继承关系、属性关系、实例关系等。科技文献知识元本体中对象属性的具体描述见表 1。除具有对称性的对象属性外，其他对象属性都具有逆属性，如引出关系的逆属性为被引出关系等，由于篇幅限制，本文不再详述。

3.2.3　科技文献知识元本体的方法集　在方法集的设计阶段，要保证本体模型的明确性、一致性，定义的各种方法要有意义且不存在矛盾。科技文献知识元本体的方法集体现了科技文献中概念、属性之间的函数

表 1　科技文献知识元本体的对象属性集

ID	对象属性名	父属性	特性	D_{PO}	属性的内涵说明
1	同位关系	无	无	0	描述同一级别概念间的语义关系
2	等级关系	无	无	0	描述不同级别概念间的语义关系
3	比…有创新	同位关系	传递性	1	概念 B 比概念 A 有创新性
4	与…同义(如阐述..)	同位关系	传递性	1	概念 B 与概念 A 同义,如概念 B 阐述了概念 A
5	位于…内/定位于..	同位关系	传递性	1	概念 A 位于概念 B 的空间范围内
6	相关	同位关系	传递性、对称性	0	描述具有一定相关性的语义关系
7	顺序	同位关系	非对称性	0	描述具有次序性的语义关系
8	与…相交于(与..合作)	相关	对称性	1	概念 A 与概念 B 具有交叉部分,如实体 A 与实体 B 具有合作关系
9	与…并列	相关	对称性	1	概念 A 与概念 B 属于同级并列关系
10	与…互斥	相关	对称性	1	概念 A 与概念 B 不存在交叉关系
11	引出	顺序	传递性	1	概念 A 引出了概念 B
12	先于(次序)	顺序	传递性	2	概念 A 排列在概念 B 之前
13	演化为	顺序	传递性	1	概念 A 可演化为概念 B
14	蕴含	顺序	传递性	1	概念 A 蕴含着概念 B(隐性关系)
15	撰写了	顺序	非对称性	1	概念 A 撰写了概念 B
16	构成了	等级关系	传递性	1	概念 A 构成了概念 B(显性关系)(部分与整体关系)
17	继承于	等级关系	传递性	1	概念 A 继承于父概念 B
18	属于	等级关系	传递性	1	属性 A 属于概念 B
19	是…的实例	等级关系	传递性	1	实例 A 是概念 B 的实例

注:对象 A 位于对象属性之前,对象 B 位于对象属性之后

关系。由于题名和摘要信息更能代表文本的主题信息[26],以此构建科技文献句子级创新点的识别方法,可抽象描述理论或方法创新点的识别过程;依据科技文献技术创新点较多出现在核心研究内容章节的规律,定义了技术创新点的识别方法;借鉴已有的学术文献文体特征研究[24],简化了文本长度的计算方法,具体内容如表 2 所示:

表 2　科技文献知识元本体的方法集

方法名	方法描述	对象类型	D_M	方法说明
理论创新点的识别法	(题名. 特征词 ∧ 主题词 ∧ 目的概述. 特征词 ∧L. 特征词) ≠ Φ→L ∈ 理论创新点	C,P	1	若某知识元 L 的特征词与该科技文献题名的特征词、主题词、结论概述中的特征词均存在相同/相似词时,则该知识元 L 就是理论创新点
方法创新点的识别法	(题名. 特征词 ∧ 主题词 ∧ 方法概述. 特征词 ∧L. 特征词) ≠ Φ→L ∈ 方法创新点	C,P	1	若某知识元 L 的特征词与该科技文献题名的特征词、主题词、方法概述中的特征词均存在相同/相似词时,则该知识元 L 就是方法创新点
技术创新点的识别法	(S. 特征词 ∧ 研究技术. 特征词) ≠ Φ ∧ (S. 特征词≠已有技术成果. 特征词)→S ∈ 技术创新点	C,P	1	若存在某知识元 S 的特征词与该科技文献中研究技术特征词相同,且知识元 S 的特征词与已有技术成果的特征词不相同,则该知识元 S 属于技术创新点
文本长度计算法	(S. 原文内容 ≠ Φ) ∧ (S. 表示形式 = 文字)→(S. 文本长度 = Length(原文内容))	P,P	2	用某知识元 S 的原文内容非空且表示形式为文字,则字符长度值即为其文本长度

注:C 代表科技文献知识元本体概念类,P 代表概念类的属性;知识元 S 属于摘要、引言、研究现状、核心研究内容、实验与评价、结论与展望六大类相关的概念子类中某一类的实例,知识元 L 属于引言/研究现状/核心研究内容/实验与评价/结论与展望五大类相关的概念子类中某一类的实例

3.2.4　科技文献知识元本体的关系集　在关系集的设计阶段,要保证本体模型具有明确性、一致性、最小本体承诺,尽量避免对同一事物做出大量的推断,满足科技文献领域知识共享与交流的需求即可。该模型通过分析科技文献中的科研思路,定义科技文献知识元本体中概念、属性及方法之间的语义关系,具体内容见表 3。其中,概念间的继承关系、部分与整体关系见图 1。

对科技文献概念进行扩展分层,限定概念的属性特征,描述概念及属性之间的方法集,表示出概念、属性及方法之间的语义关系,最终构建出一个科技文献知识元本体模型。

表 3 科技文献知识元本体的关系集

ID	对象	关系	对象	对象类别	D_R
1	形式特征	相斥于	内容特征	C,C	1
2	已有技术成果	相斥于	已有方法成果	C,C	1
3	已有方法成果	相斥于	已有理论成果	C,C	1
4	作者	撰写	书名	C,C	1
5	作者	与...合作	作者	C,C	1
6	理论创新点	比…有创新	已有理论成果	C,C	1
7	方法创新点	比…有创新	已有方法成果	C,C	1
8	技术创新点	比…有创新	已有技术成果	C,C	1
9	引言	引出	研究现状	C,C	1
10	研究现状	引出	研究内容	C,C	1
11	研究内容	引出	实验与评价	C,C	1
12	实验与评价	引出	结论与展望	C,C	1
13	概念定义	引出	研究背景	C,C	1
14	研究背景	引出	关键问题	C,C	1
15	关键问题	引出	研究目的	C,C	1
16	研究目的	引出	研究思路	C,C	1
17	已有研究成果	引出	现存问题	C,C	1
18	已有理论成果	与…并列	已有方法成果	C,C	1
19	已有方法成果	与…并列	已有技术成果	C,C	1
20	研究理论	先于	研究假设	C,C	2
21	实验数据	先于	实验工具	C,C	2
22	实验工具	先于	数据处理与分析	C,C	2
23	数据处理与分析	先于	实验结果	C,C	2
24	评价指标	先于	评价方法	C,C	2
25	评价方法	先于	评价结果	C,C	2
26	已有方法成果	演化为	研究方法	C,C	1
27	已有理论成果	演化为	研究理论	C,C	1
28	已有技术成果	演化为	研究技术	C,C	1
29	书名	蕴含	主题	C,C	1
30	分类号	蕴含	主题	C,C	1
31	分类号	蕴含	学科	C,C	1
32	文献题目	蕴含	主题	P,C	1
33	概念类. 特征词	蕴含	主题	P,C	1
34	概念类. 引文编号	位于…内	文献编号	P,C	2
35	汉语	与…相斥	外语	P,P	1
36	文本长度计算法	引出	文本长度	M,P	2
37	理论创新点识别法	引出	理论创新点	M,C	1
38	方法创新点识别法	引出	方法创新点	M,C	1
39	技术创新点识别法	引出	技术创新点	M,C	1
40	理论创新点识别法	与…并列	方法创新点识别法	M,M	1

注:C 代表科技文献知识元本体概念类,P 代表概念类的属性,M 代表概念类相关的方法

4 示例及讨论

4.1 示例

实验将以一篇科技文献为例,基于科技文献知识元本体模型,对该科技文献中的知识元进行表示与描述,用于说明上述知识表示模型的合理性。采用 protégé 4.3.0 版本的本体开发工具构建知识元本体模型,通过 protégé 自带的 HermiT 推理机实现了对该本体连续性和一致性的检测,表明该模型具有逻辑推理的能力。由于 protégé 4.3.0 版本的推理结果未能正确地显示出本体模型中概念类的中文标签,故本实验将运用英文标签呈现该本体模型的表示及推理的全过程。

4.1.1 科技文献知识元本体模型的构建 由于 CNKI、万方等学术期刊数据库已采用科技文献知识元本体模型的外部特征对文献进行分类,故本实验主要构建科技文献知识元本体模型的内容特征单元,呈现

秦春秀，杨智娟，赵捧未，等. 面向科技文献知识表示的知识元本体模型[J]. 图书情报工作,2018,62(3):94 – 103.

科技文献的内容结构。科技文献知识元本体模型的部分内容如图 2 所示：

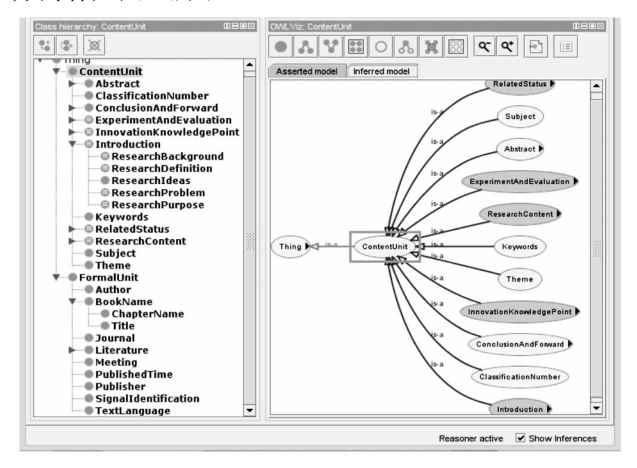


图 2　科技文献知识元本体模型(部分)

4.1.2　科技文献知识元本体模型的表示及推理　从万方数据库中，选择近两年(2016 – 2017 年)在国内情报学核心期刊《情报学报》中被引次数最高且与本文主题最相关的文献作为示例文献，即由黄永、陆伟及程齐凯撰写的题名为《学术文本的结构功能识别——基于章节内容的识别》的科技文献。该文献的写作结构规范，完整地描述出引言类中 5 个概念子类的实例知识元，具有很好的代表性。

由科技文献知识元本体的层次结构可知，引言(Introduction)知识元可细分为概念定义、研究背景、关键问题等知识元。在引言部分，一般先对某领域中主要的概念进行描述，再深入到该主题领域的研究背景，剖析出目前仍存在的亟待解决的关键问题，形成一条“概念定义(ConceptDefinition)→研究背景(Research-Background)→关键问题(ResearchProblem)”知识链。若从某篇科技文献的引言中找出概念定义的知识元 10、关键问题的知识元 12、及另一个知识元 11，而该知识元 11 在语义上是由知识元 10 引出的，进而引出知识元 12，则需判断从引言中找出的知识元 11 是否属于研究背景。将这样的逻辑思路输入 protégé 本体编辑器中，运行 HermiT 推理机可得，知识元 11 属于引言中的研究背景知识元。该结果在 Description 11 的 Type 类中的虚线框内，即表明上述逻辑关系得到了验证，详细的推理结果见图 3。

另外，在对科技文献知识元本体模型的方法集进行演绎时，科技文献中理论创新点的识别方法为：待判定知识元的特征词与题名(Title)、主题词(Theme)、及摘要中目的概述(PurposeOverview)的特征词均有交集，则该知识点是理论创新点(TheoryInnovation)。从核心研究



图 3　科技文献知识元本体模型的表示及推理

内容中找到某一理论知识点(KnowledgePoint01)，依据上述判定规则，分别创建题名、主题词、目的概述的实例 01。由于待判定的知识点实例的特征词含有题名实例 01(Title01)、主题词实例(Theme01)和目的概述实例(PurposeOverview01)中的同一特征词，验证出该知识点是文献的理论创新点。具体的操作痕迹见图 3。

4.2　分析与讨论

笔者通过分析科技文献的文本结构，在知识元本体理论的基础上，构建一个科技文献知识元本体模型，用于表示科技文献中知识元的结构和知识元之间的语义关系。该知识元本体模型在内容呈现、构建方法、应用场景等方面均呈现较好的效果，具体包括以下几点：

(1)呈现清晰的知识组织体系。笔者构建的科技文献知识元本体模型具备横向和纵向双重组织结构：①在该模型的横向结构中，从科技文献的外部特征和内容特征着手，逐层深入科技文献的章节、段落中，搜寻科技文献中的功能性知识点，最终呈现出科技文献内不同层次的知识结构；②在该模型的纵向结构中，从科研习惯的角度，首先关注科技文献显而易见的外部特征，对科技文献的作者、文献类型、出版社、出版时间等作初步筛选，再深入科技文献内容中，阅读摘要、引言等章节，快速了解该文献的大致内容，如果发现有用户关注的相关主题知识，再继续深入挖掘该章节中用户所需的精准知识元。故该本体模型能清晰地呈现科技文献领域中的文本结构及知识脉络。

(2)更为精准、高效的模型构建法。该知识元本

体模型的构建法融合了本体构建法和知识元构建法的双重优势：①在本体构建方法的指引下，构建出的本体概念框架能抽象化描述科技文献中通用的文本结构，据此可将篇章级的科技文献分割为句子级的细粒度知识点。另外，依据所构建的本体模型中的知识链网络，不仅能精准定位并表示科技文献中丰富的知识点，而且能有效地推动科技文献内知识的推理及发现。②融入知识元的 n 元组构建法的思想，对科技文献中知识元所属概念类的属性特征进行规范，采用相对统一的属性集，简洁、高效地表示出科技文献中摘要、引言、研究现状等知识元以及研究背景、研究问题、研究方法等更细粒度知识元的内部结构，有利于学术论文、期刊论文、学位论文、专利等各类文献资料中知识的表示、存储、共享与交流。

（3）具备广泛的应用前景。该知识元本体模型适用于知识表示、知识组织和知识标引场景中，辅助不同层次的用户学习不同领域的科研知识：①在知识表示的应用场景中，可依据该本体模型，对某领域科技文献中的重要知识点进行描述和标注，便于用户快速了解本领域科技文献的知识结构，辅助具有其他学科背景的用户快速学习本领域的专业知识；②在知识组织的应用场景中，对于还未形成规范的科研及写作思路的学术研究者，可依据该模型中呈现的科技文献文本结构，了解本领域科技文献的科研思路及写作模式，缩短高水平科研成果的产出周期，促进多领域学科的快速发展；③在知识标引的应用场景中，该本体模型可被应用于在线投稿系统。在论文投稿时，编辑部将依照该本体模型，设定投稿论文应具备的主要知识模块，规范投稿论文的内容及格式；当该论文已被录用后，编辑部应鼓励并引导该论文作者以该本体模型的知识结构为标准，对论文内容进行精准分解及标注，促进科技文献领域知识的协同标引。

综上所述，该本体模型能较好地表示和描述科技文献的文本结构及知识元的内部结构关联，拥有广泛的应用前景，且能为广大用户提供更加高效的文献知识服务。

5 结论与展望

为了提供细粒度的精准知识点服务，笔者在分析科技文献内容的层次性功能结构的基础上，提出一种面向科技文献的知识表示模型，并以一篇科技文献为例，统一表示科技文献中细粒度知识元的内外部特征及知识元之间的逻辑关系，为多角度、精准、快速的知识点检索奠定坚实的理论基础。该本体模型可作为科技文献内知识资源的表示及存储模式标准，用于改善现有粗粒度的文献知识服务模式，获取粒度更小、层次更深的科技资源。将该本体模型应用于科技文献知识资源的组织与协同标引场景中，将会改变论文出版商和科研工作者的知识发布、使用、传播、获取方式，促进细粒度知识资源的共享及创新知识点的发现。

笔者仅运用一个示例还不足以验证该模型的有效性，该模型仍存在以下几点不足：①科技文献知识元本体模型仅呈现了科技文献领域通用的文本结构，而在不同学科、主题等领域下，科技文献的文本结构会有所不同，仍需要进一步细化该模型的适用领域，完善该模型的本体结构；②该本体模型仍未得到科学的评估及验证，在今后的工作中将会深入研究科技文献知识元本体模型的评价方法。

参考文献：

[1] GU X, BLACKMORE K L. Recent trends in academic journal growth[J]. Scientometrics,2016,108(2): 693 – 716.

[2] 化柏林，刘一宁，郑彦宁. 针对学术定义的抽取规则构建方法研究[J]. 情报理论与实践,2011,34(12):5 – 9.

[3] 化柏林. 学术论文中方法知识元的类型与描述规则研究[J]. 中国图书馆学报,2016,42(1):30 – 40.

[4] 温有奎，温浩，徐端颐，等. 基于创新点的知识元挖掘[J]. 情报学报,2005,24(6):663 – 668.

[5] 秦春秀，刘杰，刘怀亮，等. 基于知识元的科技文本内容描述框架研究[J]. 图书情报工作, 2017,61(10):116 – 124.

[6] 肖泉，蔡淑琴，叶波. 基于超图结构的知识相似度计算模型研究[J]. 情报学报,2010,29(5):805 – 812.

[7] 马创新，陈小荷，曲维光. 经典古籍注疏文献的知识网络研究与设计[J]. 图书情报工作,2013,57(9):124 – 128.

[8] 吴思竹，李峰，张智雄. 知识资源的语义表示和出版模式研究——以 Nanopublication 为例[J]. 中国图书馆学报,2013,39(4):102 – 109.

[9] 王宇，刘淼. 一种基于知识元的期刊文献知识仓库构建[J]. 情报理论与实践,2013,36(8):91 – 94.

[10] 索传军. 知识转移视角下的学术论文老化与创新研究[J]. 图书情报工作,2014,58(5):5 – 12.

[11] 高俊芳. 云计算下的高校图书馆学科知识服务研究[J]. 图书馆学研究,2015,33(2):54 – 58.

[12] 化柏林. 基于 NLP 的知识抽取系统架构研究[J]. 现代图书情报技术,2007,2(10):38 – 41.

[13] 宋继华，李国玉，王宁.《说文解字》音义关系的产生式表达[J]. 中文信息学报,2006,20(2):53 – 59.

[14] 徐晨飞，倪媛，钱智勇. 基于本体的“江海文化”文献知识组织体系构建研究[J]. 现代情报,2015,35(10):62 – 71.

[15] MARCONDES C H, COSTA L C D. A model to represent and process scientific knowledge in biomedical articles with semantic

web technologies[J]. Knowledge organization,2016,43(2):86－101.
[16] 熊李艳,陈建军,钟茂生. 基于 E-A-V 结构的概念图匹配算法[J].计算机应用研究,2014,31(8):2290－2293.
[17] 李军锋,吕学强,李卓. 专利领域本体概念语义层次获取[J].情报学报,2014(9):986－993.
[18] 温有奎,焦玉英. Wiki 知识元语义图研究[J].情报学报,2009,28(6):870－877.
[19] 姜永常. 知识网络链接的理论基础与基本原则[J].图书馆,2012(2):31－34.
[20] 魏伟,郭崇慧,唐琳,等. 基于知识元的文献挖掘研究——以粤海关文献资料为例[J].情报科学,2017(6):138－144.
[21] 王静,刘成山,秦春秀. 一种基于模糊 Petri 网的知识元语义集成方法[J].情报理论与实践,2017,40(9):140－144.
[22] 温浩,温有奎. 主题成因的知识元本体转换模型研究[J].情报学报,2011,30(11):1123－1128.
[23] 温有奎,徐国华."认知元"的三维结构理论[J].情报学报,2004,23(2):242－246.
[24] 邹永利. 学术文献的非主题特征及其意义[J].中国图书馆学报,2011,37(3):100－107.
[25] 焦玉英,符绍宏,何绍华. 信息检索[M].武汉:武汉大学出版社,2008.
[26] 李湘东,丁丛,高凡. 基于复合加权 LDA 模型的书目信息分类方法研究[J].情报学报,2017,36(4):352－360.
[27] VAUGHAN M W, DILLON A. The role of genre in shaping our understanding of digital documents[EB/OL].[2017－11－15]. http://arizona.openrepository.com/arizona/bitstream/10150/105666/1/MvAd1998.pdf.
[28] DING Y, LIU X, GUO C, et al. The distribution of references across texts: some implications for citation analysis[J]. Journal of informetrics,2013,7(3):583－592.
[29] 王鹏,赵逢禹,陈章. 基于分层分割的科研领域文本信息挖掘[J].情报学报,2015(1):85－91.
[30] YEPES A J, MORK J, ARONSON A. Using the argumentative structure of scientific literature to improve information access[EB/OL].[2017－11－15]. http://www.aclweb.org/old_anthology/W/W13/W13-19.pdf#page=114.
[31] 朱丽萍,李洪奇,杨中国,等. 一种面向科技文献引言的信息抽取方法[J].山东大学学报理学版,2015,50(7):23－30.
[32] 陆伟,黄永,程齐凯. 学术文本的结构功能识别功能框架及基于章节标题的识别[J].情报学报,2014(9):979－985.
[33] 化柏林. 针对中文学术文献的情报方法术语抽取[J].现代图书情报技术,2013(6):68－75.
[34] 连燕华,马晓光. 评价过程分析模型探讨[J].研究与发展管理,1999(3):1－5.
[35] ZHANG L. Grasping the structure of journal articles: utilizing the functions of information units[J]. Journal of the Association for Information Science & Technology,2012,63(3):469－480.
[36] GRUBER T R. Toward principles for the design of ontologies used for knowledge sharing[J]. International journal of human computer studies,1995,43(5－6):907－928.
[37] 王洪伟,霍佳震,王伟,等. 面向语义检索应用的本体模型结构设计[J].系统工程与电子技术,2010,32(1):166－174.
[38] 张晗,徐硕,乔晓东. 融合科技文献内外部特征的主题模型发展综述[J].情报学报,2014(10):1108－1120.

作者贡献说明:
秦春秀:提出研究问题及思路,给出修改意见;
杨智娟:设计研究方案,撰写并修改论文;
赵捧未:提出修改意见;
刘杰:提出修改意见。

The Knowledge Element Ontology Model of Scientific Literature for Knowledge Representation

Qin Chunxiu　Yang Zhijuan　Zhao Pengwei　Liu Jie
School of Economics and Management, Xidian University, Xi'an 710071

Abstract: [**Purpose/significance**] With the rapid growth of scientific literature resources, users are surrounded by the ocean of scientific literature, and the future trend of the document knowledge retrieval is to provide users with fast and accurate fine-grained knowledge element services. [**Method/process**] Based on the analysis of the content structure of scientific literature, this paper went gradually and deeply into the scientific literature contents, with a view of constructing a knowledge element ontology model of scientific literature for the knowledge representation, to express fine-grained knowledge points owning holonomic sentence meanings in the content of scientific literature as knowledge elements with a unified structure. [**Result/conclusion**] The paper illustrates the rationality of the model of scientific literature by the means of displaying the content of a scientific literature, but this example only shows knowledge points in the introduction of scientific literature and the knowledge element ontology model needs to be further verified.

Keywords: scientific literature　knowledge representation　knowledge element　ontology model